

基于对象频度的粗集约简算法研究

朱万富,李 坤

(安徽理工大学 机械工程学院,安徽 淮南 232001)

摘 要:决策表中大量的不相容对象和相同对象会降低粗集约简算法性能,出现错误的属性核和约简结果。因此文中首先讨论决策表的不相容对象与相同对象即冗余对象对约简结果的影响,然后提出利用对象频度概念来删除冗余对象,组建最简决策表的算法。把求最简决策表的算法与高学东的粗集约简算法结合起来,形成了一种具有较好的处理不相容决策表、决策规则较合理的粗集约简算法。实例证明,该算法性能时间与空间复杂度与高学东算法相同,但明显优于 Hu Xiaohua、叶东毅等的算法。

关键词:粗集;最简决策表;对象频度;约简;算法

中图分类号:TP301.6

文献标识码:A

文章编号:1673-629X(2014)03-0039-03

doi:10.3969/j.issn.1673-629X.2014.03.010

Research on Rough Set Reduction Algorithm Based on Object Frequency

ZHU Wan-fu, LI Kun

(School of Mechanical Engineering, Anhui University of Science and Technology, Huainan 232001, China)

Abstract: Incompatible objects and the same objects in the decision table can degrade the performance of rough set reduction algorithms, producing wrong attribute core and reduction results. Therefore firstly it is discussed that incompatible objects and the same objects also called redundancy objects of the decision table affect reduction results. Secondly, objects frequency is put forward to remove redundant objects, the simplest decision table algorithm has been proposed. The algorithm for the simplest decision table is integrated with Gao Xuedong's rough reduction algorithm, which could better deal with inconsistent decision tables and decision rules is more reasonable. Example proves that the performance time and space complexity of this algorithm is same as Gao Xuedong's algorithm, and is better than Hu Xiaohua's, Ye Dongyi's algorithms.

Key words: rough set; simplest decision table; object frequency; reduction; algorithm

0 引言

波兰数学家 Skowron 提出的基于差别矩阵的约简方法^[1],是众多粗集约简算法中较优秀的一种—易实现且原理简单,但它也有占有空间与时间太大、性能较差的缺点,因此国内外许多学者对其进行改进,如刘清等提出可辨矩阵方法^[2-3],乐艳丽等提出启发式信息为知识量的约简算法^[4],苗夺谦等提出相异矩阵方法^[5-6],冯林等人提出创建多变量决策树进行约简的算法^[7-8],高学东提出改进差别矩阵,组成简化决策表后用区分对象对集进行约简,算法性能得以提高^[9]。

以上这些基于差别矩阵的算法,均只考虑改进差别矩阵来提升算法的性能,却忽略了决策表中大量的不相容对象和相同对象,这些对象会降低约简算法性能,出现错误的属性核和约简结果。因此文中首先讨

论决策表的不相容对象、相同对象对约简的影响,然后提出基于频度的粗集约简算法。

1 基本概念定义

为了方便,文中首先定义讨论所涉及到的概念:

定义1:不相容对象。决策表中任意两对象 $x, y \in U$, 当 $x \neq y$ 时,如果 $f(x, c) = f(y, c)$ 可推出结论 $f(x, d) \neq f(y, d)$, 则对象 x 与对象 y 是不相容对象^[10]。

定义2:相同对象。决策表中存在着两对象 $x, y \in U$, 当 $x \neq y$ 时,如果由 $f(x, c) = f(y, c)$ 可推出结论 $f(x, d) = f(y, d)$, 则对象 x 与对象 y 是相同的对象^[10]。

定义3:频度 $g(x)$ 。频度 $g(x)$ 就是对象 x 在决策表中出现的次数^[10]。

定义4:冗余对象。决策表中可删除的相同对象

收稿日期:2013-05-29

修回日期:2013-09-07

网络出版时间:2014-01-07

基金项目:安徽省高等学校优秀青年人才基金重点资助项目(2012SQRL044ZD);安徽理工大学引进人才基金项目

作者简介:朱万富(1974-),男,安徽人,硕士,研究方向为人工智能、机械故障诊断。

网络出版地址: <http://www.cnki.net/kcms/detail/61.1450.TP.20140107.1648.030.html>

和不相容对象统称为冗余对象^[10]。

定义 5: 最简决策表。如果决策表 $S = (U, C \cup D)$, 对象 $x, y \in U$, 如果当 $x \neq y$ 时, 对任意的 $B \subseteq C$, 不存在 $f(x, B) = f(y, B)$, 则该决策表 S 为最简决策表^[10]。

定义 6: 冗余决策表。存在着冗余对象的决策表即为冗余决策表。冗余决策表中存在着可删除的冗余对象。决策表中仅存在相同对象时, 称为相容决策表; 决策表中存在不相容对象时, 称为不相容决策表^[10]。

相容决策表中, 条件属性取值相同的对象其决策属性值也必定相同; 不相容决策表中, 至少存在条件属性取值相同的两个对象, 但决策值却不相等; 而冗余决策表中, 至少存在条件属性取值相同的两个对象, 决策属性值可能相同, 也可能不同^[10]。

由最简决策表定义可以得知: 最简决策表等价于冗余决策表, 其实最简决策表的任一元素一定与冗余决策表的一元素对应, 这两种决策表拥有相同的信息量, 形成决策规则也一样, 但最简决策表中没有冗余对象, 约简算法性能可以大大提高, 因此文中讨论怎样得到最简决策表^[10]。

2 现有粗集约简算法存在的问题

决策表中冗余对象对约简主要有以下影响:

(1) 求得的属性核和约简可能是错误的, 由于约简算法没有充分考虑决策表不相容性及其程度的影响。例如表 1。

表 1 决策表 1

	<i>a</i>	<i>b</i>	<i>c</i>	<i>D</i>
M_1	1	1	0	1
M_2	1	1	0	0
M_3	1	0	0	1
M_4	1	0	0	2
M_5	1	0	0	2
M_6	1	0	0	2

a, b, c 为条件属性, D 为决策属性, 对象为 $M_1 \sim M_6$, 不相容对象有 M_1 与 M_2, M_3, M_4, M_5 与 M_6 , 该决策表为冗余决策表, 基于差别矩阵的约简算法得到的核属性为 $\{a, b, c\}$, 但真实情况是 b 不是核, 约简结果是错误的^[10]。

(2) 决策表中冗余对象若不消除, 将使约简算法重复计算, 导致算法的性能降低, 受到影响。如某决策表中有对象 $M_1 \sim M_{14}$, 但不相容对象有: M_1 与 M_3, M_2, M_6 与 M_9, M_5 与 M_{10} , 相同对象有: M_4 与 M_8, M_5 与 M_7, M_{11} 与 M_{14} , 因此该表的冗余程度较大。粗集约简算法若对这些对象一视同仁, 进行约简, 将在这些冗余对象

上进行不必要的时间开销, 因此算法的性能较低^[10]。

(3) 算法如选择不相容对象不当, 得到的决策规则有可能是错误的。

冗余决策表中存在着大量不相容对象, 尹一麒等^[11]认为不一致的对象对决策规则的形成不起作用, 这种看法值得探讨: 如考虑决策系统完备性, 则必须考虑不相容对象的作用, 但选取不相容对象中哪一条对象来形成决策规则, 则影响到整个决策系统的组成。目前的约简算法很少考虑这个问题, 因此选取不相容对象时具有相当大的随意性, 如高学东的约简算法其实选取了不相容对象中的第一个对象形成决策规则, 却忽略了不相容对象中相同对象的频度对决策的影响因素。例如某决策表中有 20 个不相容对象-条件属性值相同而决策属性值不同, 第 1 个对象的决策属性值为一个值, 后面的 19 个对象决策属性值相同为另一个值, 很明显后面这相同的 19 个对象对决策的影响应大于第 1 条, 应选取入决策系统, 但按高学东算法, 却选取第 1 个对象进入决策系统, 因此以后可能形成错误的决策规则^[10]。

3 基于对象频度的求最简决策表算法

文中提出删除冗余对象, 组成最简决策表的方法: 相同对象产生相同的决策规则, 决策表中只须保留这些相同对象中的任一对象即可; 不相容对象, 算法将根据该对象的频度-即选取频度数最大的不相容对象来构建最简决策表, 形成决策规则, 这点不同于高学东的算法中暗含的选取最先出现的不相容对象选择规则。

因此文中基于以上的思想, 改进高学东求 U/C 的算法^[9] 后得到了基于对象频度的求最简决策表的算法, 提出的算法步骤如下:

算法输入: 冗余决策表 $S = (U, C, D, V, f, d)$ 初始决策表; $U = \{x_1, x_2, \dots, x_n\}$ 为对象集合, 共 n 个对象; $C = \{C_1, C_2, \dots, C_s\}$ 为属性集合, 共 s 个属性。

算法输出: 最简决策表 $S', U/C$ 。

第一步: 令 $SET[0] = \{U\}$;

第二步: for ($i = 1; i < s + 1; i++$)

{

SET[i] = \emptyset ; $K = 1$;

while (SET[$i - 1$]! = \emptyset)

{

取出 SET[$i - 1$] 的第一个元素 S ;

SET[$i - 1$] = SET[$i - 1$] - $\{S\}$;

while ($S!$ = \emptyset) // 以下均为对 S 的操作

{

$H_{ik} = \emptyset$;

S 中的第一个元素 x 并入 H_{ik} 并在 S 中删除元素 x ;

对 S 中的其他元素 y :

```
if  $f(y, c_i) = f(x, c_j)$ 
{ 将元素  $y$  并入  $H_{jk}$ ; 在  $S$  中删除元素  $y$  }
SET[ $i$ ] = SET[ $i$ ]  $\cup$  { $H_{ik}$ };
 $k = k + 1$ ;
}
}
}
第三步: 设第二步得到的集合 SET[ $S$ ] = { $H_{s1}, H_{s2}, \dots, H_{st}$ };
for ( $i = 1; i < t + 1; i++$ )
while ( $H_{si} \neq \emptyset$ )
{
依次取出  $H_{si}$  中的每一个对象  $x$ ;
if ( $g(x) \neq 0$ )
{
 $g(x) = 1$ ; 取出  $H_{si}$  中的其他任一对象  $y$ 
if ( $f(x, d) = f(y, d)$ )
{  $g(x) = g(x) + 1; g(y) = 0$  }
}
}
第四步: 求解最简决策表与  $U/C$ 。
 $U'_{pos} = \emptyset; U'_{neg} = \emptyset; U' = \emptyset$ ;
for ( $i = 1; i < t + 1; i++$ )
依次取出  $H_{si}$  中的对象  $x$  的  $g(x)$  值
if ( $H_{si}$  中所有对象仅有一个  $g(x) \neq 0$ )
{ 将此对象  $x$  插入最简决策表并入  $U'_{pos}$  中; 删除  $H_{si}$  中的其他对象 }
else if ( $g(x)$  为  $H_{si}$  中的  $g(x)$  的最大值)
{ 将此对象  $x$  插入最简决策表与  $U'_{neg}$  中 }
else
{ 删除此对象 }
```

算法与高学东算法有相同的复杂度, 空间复杂度为 $O(|C| + |U|)$, 时间复杂度为 $O(\sum_{i=1}^{|C|} |k_i| + |U|)$ (其中 $|k_i|$ 表示第 i 个属性的不同取值个数)。算法优于高学东算法之处在于: 原决策表中大量的冗余对象被删除, 频度最大的不相容对象被保留组成最简决策表, 因此得到的决策规则是完备的^[10]。

4 基于对象频度的粗集约简算法

基于对象频度的粗集约简算法原理为: 在决策表约简时, 首先利用上节得到的基于对象频度的求最简决策表算法删除全部冗余对象, 得到最简决策表, 再使用高学东算法中基于区分对象对集的约简算法对最简决策表进行约简, 这样既保证了算法的高效性又保证了算法结果的正确性。

基于对象频度的粗集约简算法具有高效性, 经分析得知, 时间复杂度与高学东算法相同, 为 $\max\{O(|C| + |U'_{pos}| + |U/C|), O(|C| + |U|)\}$, 空间复杂度

为 $\max\{O(|C| + |U'_{pos}| + |U/C|), O(|C| + |U|)\}$, 明显比 Hu XiaoHua、叶东毅等提出的性能均为 $O(|C|^2 + |U|^2)$ 的算法优越^[12-13]; 优越之处为: 具有较好的处理不相容决策表的能力, 不管是相容决策表还是不相容决策表, 以及表中对象冗余程度有多大, 均能进行约简并且得出结果正确, 形成决策规则较合理的决策系统。

5 算法实例

下面以表 1 决策表为例, 分别使用基于对象频度的粗集约简算法与高学东算法对决策表进行约简。基于对象频度的粗集约简算法得到的决策表为表 2, 高学东算法约简后得到的决策表为表 3。

对比表 2 与表 3, 发现两表的区别在于: 最简决策表选取不相容对象组中的 M_4 形成决策规则, 而高的算法中选取了 M_3 。原表中 M_4 的对象频度为 3, 而 M_3 对象频度为 1, 因此文中算法选取 M_4 形成决策规则较为合理。

表 2 最简决策表				
	a	b	c	D
M_1	1	1	0	1
M_2	1	1	0	0
M_3	1	0	0	1

表 3 简化的决策表				
	a	b	c	D
M_1	1	1	0	1
M_2	1	1	0	0
M_4	1	0	0	2

6 结束语

文中首先讨论了决策表中冗余对象对约简的影响, 提出一种基于对象频度的求最简决策表算法—利用对象频度的概念, 删除频度低的不相容对象和相同的对象, 组建最简决策表, 然后把求最简决策表的算法与高学东约简算法结合起来, 形成了一种粗集约简算法—基于对象频度的粗集约简算法。它具有较好的处理不相容决策表的能力, 能够形成决策规则较合理的决策系统, 算法性能如时间与空间复杂度与高学东算法相同, 但明显优于 Hu XiaoHua、叶东毅等的算法。

参考文献:

[1] Skowron A, Crauszer. The discernibility matrix and functions in information system [M]//Handbook of applications and advances of the rough set theory. [s. l.]: Kluwer Academic (下转第 45 页)

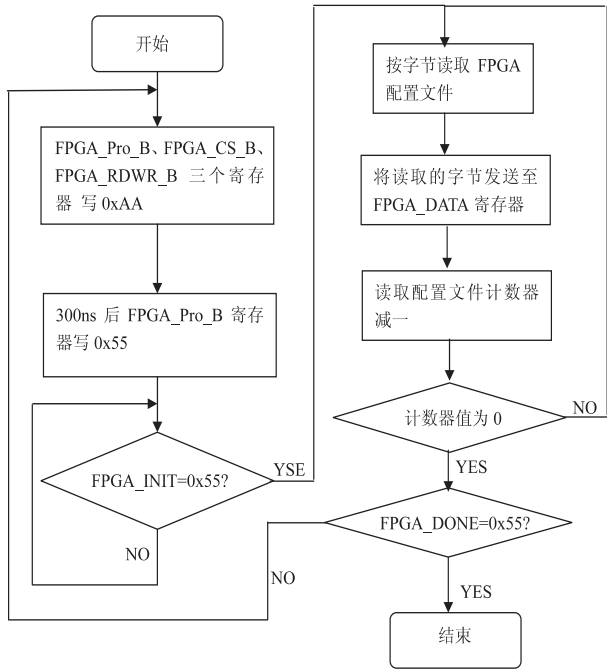


图4 DSP 配置 FPGA 软件流程图

3 结束语

文中描述了通过图像处理系统中 TMS320C6416T DSP、S29JL064H FLASH 和 XC95144XL CPLD 实现了对 Xilinx 公司 XC5VLX50 FPGA 的动态配置,文中所述方法除了可以利用现有资源实现 FPGA 上电配置,节省 FPGA 专用配置芯片,简化电路,另外可以实现 FPGA 动态配置,可以在 FLASH 内存存储几种 FPGA 配置文件,根据实际需求进行上电配置,还可以利用系统的

串口或网络接口,实现 FPGA 程序的动态更新,增加系统的灵活性和实用性。

参考文献:

[1] Peattie M. Using a microprocessor to configure Xilinx FPGAs via slave serial or SelectMAP mode XAPP502 [R]. [s.l.]: [s.n.],2009.

[2] Xilinx. Virtex-5 FPGA configuration user guide [R]. [s.l.]: Xilinx,2009.

[3] Xilinx. Virtex-5 FPGA user guide [R]. [s.l.]: Xilinx,2009.

[4] Xilinx. Virtex-5 FPGA data sheet:DC and switching characteristics [R]. [s.l.]: Xilinx,2009.

[5] Xilinx. Xilinx in-system programming using an embedded microcontroller [R]. [s.l.]: Xilinx,2009.

[6] Xilinx. Virtex-4 FPGA configuration user guide [R]. [s.l.]: Xilinx,2009.

[7] 余佑军,王 丹. TMS320C61416 EMIF 总线下双 FPGA 加载设计[J]. 单片机与嵌入式系统应用,2007(7):29-31.

[8] 王 雄,靳济方. 一种 FPGA 动态配置的实现方法[J]. 北京电子科技学院学报,2007,15(2):46-48.

[9] 贾 鹏,李 松. FPGA 的动态可重配置技术[J]. 军事通信技术,2001,22(2):52-54.

[10] 李 晶,钟 瑜,郑百衡. 基于 DSP 的 FPGA 动态配置技术[J]. 电讯技术,2005,45(1):156-159.

[11] 范 斌,常 青. 基于 DSP 的 FPGA 动态重构系统研究与设计[J]. 信息与电子工程,2010,8(2):123-127.

[12] 谷 鑫,徐贵力,王友仁. FPGA 动态可重构理论及其研究进展[J]. 计算机测量与控制,2007,15(11):1415-1418.

(上接第 41 页)

Publishers,1999:331-361.

[2] 刘 清. Rough 集及 Rough 推理[M]. 北京:科学出版社,2001.

[3] 刘 清,刘少辉,郑 非. Rough 逻辑及其在数据约简中的应用[J]. 软件学报,2001,12(3):415-419.

[4] 乐艳丽,鲁汉榕,高 婷. 基于粗集理论的属性约简改进算法[J]. 计算机工程与设计,2008,29(13):3432-3434.

[5] 苗夺谦,胡桂荣. 知识约简的一种启发式算法[J]. 计算机研究与发展,1999,36(6):681-684.

[6] 王 钰,王 任,苗夺谦,等. 基于 RoughSet 理论的“数据浓缩”[J]. 计算机学报,1998,21(5):393-400.

[7] 冯 林,刘照鹏,黄志伟. 基于粗集理论的决策表知识获取方法研究[J]. 计算机工程与设计,2006,27(12):2172-

2174.

[8] 陈世联,罗秋瑾. 基于粗集和距离函数的决策树构造方法[J]. 计算机工程与设计,2008,29(12):3191-3193.

[9] 高学东,丁 军. 基于简化差别矩阵的属性约简算法[J]. 系统工程理论与实践,2006,28(6):101-107.

[10] 朱万富. 基于粗集神经网络的故障诊断专家系统研究[D]. 北京:中国石油大学,2008.

[11] 尹一麒,苗夺谦,李道国. 分体策略在差别矩阵优化中的应用[J]. 小型微型计算机系统,2007,28(2):292-296.

[12] 叶东毅,陈昭炯. 一个新的差别矩阵及其求核方法[J]. 电子学报,2002,30(7):1086-1088.

[13] Hu Xiaohua, Cercon N. Learning in relational data-base a rough set approach [J]. Computation intelligence,1995(2):323-337.

基于对象频度的粗集约简算法研究

作者：[朱万富](#)，[李坤](#)，[ZHU Wan-fu](#)，[LI Kun](#)
作者单位：[安徽理工大学 机械工程学院, 安徽 淮南, 232001](#)
刊名：[计算机技术与发展](#)

英文刊名：[Computer Technology and Development](#)

年，卷(期)：2014(3)

本文链接：http://d.g.wanfangdata.com.cn/Periodical_wjfz201403010.aspx